

SOME COMMENTS ON PRECONDITIONING FOR NORMAL EQUATIONS AND LEAST SQUARES

ANDY WATHEN*

Abstract. The solution of systems of linear(ized) equations lies at the heart of many problems in Scientific Computing. In particular for large systems, iterative methods are a primary approach. For many symmetric (or self-adjoint) systems, there are effective solution methods based on the Conjugate Gradient method (for definite problems) or MINRES (for indefinite problems) in combination with an appropriate preconditioner, which is required in almost all cases.

For nonsymmetric systems there are two principal lines of attack: the use of a nonsymmetric iterative method such as GMRES, or transformation into a symmetric problem via the normal equations and application of LSQR. In either case, an appropriate preconditioner is generally required.

We consider the possibilities here, particularly the idea of preconditioning the normal equations via approximations to the original nonsymmetric matrix. We highlight dangers that readily arise in this approach.

Our comments also apply in the context of linear least squares problems.

Key words. Preconditioning, normal equations, matrix squaring, least squares

AMS subject classifications. 65F08, 65F10, 65F20

1. Introduction. For a linear system of equations

$$(1.1) \quad Bx = b, \quad B \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n,$$

there are many methods for computing the solution $x \in \mathbb{R}^n$ when B is invertible. If n is large, the practical possibilities diminish and iterative methods often provide the only effective way. For some problems, stationary iterative methods are sufficient, but more common are Krylov subspace methods [15, 20]. Convergence rate is always a key consideration. The main approach to achieving acceptably fast convergence is through preconditioning [21]: employing an approximating matrix or linear operator P for which systems $Pz = r$ are much more readily solved than (1.1) and for which an appropriate Krylov subspace method will converge quickly on systems

$$P^{-1}Bx = P^{-1}b \quad \text{or} \quad BP^{-1}y = b, \quad x = P^{-1}y$$

(left-preconditioning and right-preconditioning respectively).

If B is symmetric, the iterative methods of choice are the Conjugate Gradient method (CG) [8] for definite systems and MINRES [12] for indefinite systems. Reliable convergence theory based only on eigenvalues exists for both methods and gives criteria for preconditioner design/choice. Several practical approaches result with reliable guarantees [1, 21, 5]. Symmetry can be preserved provided P is symmetric. For indefinite symmetric B , P must generally be definite [4].

For nonsymmetric B , the most widely employed iterative method is GMRES [16], though many other Krylov subspace methods (BiCGSTAB [19], QMR [6], IDR [18], ...) are used in different application areas. This array of possibilities indicates an important structural issue: there is no method of choice for all nonsymmetric systems—a profound issue elegantly pointed out by Nachtigal, Reddy and Trefethen [11]. Further, no generally descriptive convergence theory is yet known even for GMRES. For the other methods the convergence theory is almost non-existent. This is a significant

*Mathematical Institute, Oxford University, UK (wathen@maths.ox.ac.uk)

disadvantage because, without theory to guide us, preconditioning must generally be heuristic. An always available alternative is to work with the normal equations

$$(1.2) \quad B^T Bx = B^T b.$$

This brings back the advantage of a symmetric system, but the job of identifying appropriate preconditioners is then more complicated because one usually does not want to compute $B^T B$, which may be dense even if B is sparse. A first thought might be to approximate B by P and then employ $P^T P$ as a preconditioner for $B^T B$. Unfortunately—as deserves to be much more widely known— P can be an excellent preconditioner for B even in the symmetric case, whereas $P^T P$ can be arbitrarily poor as a preconditioner for $B^T B$. This was pointed out by Braess and Peisker—who called it the *matrix squaring problem*—in 1986 [2], but it still seems to be not at all widely appreciated.

For linear least squares problems,

$$\min_x \|b - Bx\|_2, \quad B \in \mathbb{R}^{m \times n}, m > n,$$

also solved by (1.2), we show that a similar issue can arise. In this setting, a right-preconditioner $R \in \mathbb{R}^{n \times n}$ is typically employed so that the preconditioned problem

$$y = \operatorname{argmin} \|b - BR^{-1}y\|, \quad x = R^{-1}y$$

is solved, usually by use of LSQR[13], which is mathematically equivalent to applying CG to the (preconditioned) normal equations. The preconditioner R is typically derived from some form of inexact QR factorisation of B , though some authors employ incomplete Cholesky factorisations (see, for example, [17]).

Our aim is to explore this issue of preconditioning for the normal equations and for least squares problems. We review the work of Braess and Peisker and other previous contributions, giving a number of examples that highlight the matrix squaring problem. There are situations where we fortunately need have no concerns about matrix squaring, that is, situations where fast convergence for the preconditioned normal equations with preconditioner $P^T P$ can be expected as long as P satisfies certain conditions; we review these.

1.1. Iterative methods. We need reference to stationary (also called *simple*) iterations

$$(1.3) \quad Px_{k+1} = (P - B)x_k + b, \quad k = 0, 1, \dots \quad \text{with } x_0 \text{ arbitrary}$$

associated with (1.1) for which convergence to the solution x of the sequence of iterates x_k occurs if and only if the eigenvalue spectrum $\sigma(I - P^{-1}B)$ is contained strictly inside the unit circle, that is, all eigenvalues of the iteration matrix $I - P^{-1}B$ lie strictly inside the unit circle. For a convergent iteration, the error ultimately contracts at a rate given by the absolute value of the largest eigenvalue: in any subordinate norm, for large enough k ,

$$\|x - x_{k+1}\| \leq |\lambda_{\max}(I - P^{-1}B)| \|x - x_k\| = |\lambda_{\max}(I - BP^{-1})| \|x - x_k\|,$$

because of an obvious similarity transform involving P . If $\|I - P^{-1}B\| < 1$ then certainly the iteration converges; indeed the error must reduce at every iteration. Thus, if $\|I - P^{-1}B\| = \gamma < 1$ then at every iteration k ,

$$\|x - x_{k+1}\| \leq \gamma \|x - x_k\|,$$

that is, the error must *contract* by at least γ . Associated with (1.1) is naturally (but often hidden and unnecessary) an adjoint linear system

$$(1.4) \quad B^T y = c$$

for which contraction of the corresponding stationary iteration

$$(1.5) \quad P^T y_{k+1} = (P^T - B^T) y_k + c, \quad k = 0, 1, \dots \quad \text{with } y_0 \text{ arbitrary}$$

will occur at every iteration if $\|I - P^{-T} B^T\| = \widehat{\gamma} < 1$.

We also require the common and usually descriptive convergence bound for CG applied to symmetric positive definite (SPD) systems with an SPD preconditioner. This bound applies to the normal equations (1.2) with preconditioner $P^T P$: the CG iterates x_k satisfy

$$(1.6) \quad \|x - x_k\|_{B^T B} \leq 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k \|x - x_0\|_{B^T B},$$

where $\kappa = \lambda_{\max}((P^T P)^{-1} B^T B) / \lambda_{\min}((P^T P)^{-1} B^T B)$ is the (2-norm) condition number of the preconditioned normal equations and $\|s\|_H^2 = s^T H s$ (see for example [4, Chapter 2]). Clearly there is good (fast) convergence of CG for the preconditioned normal equations (and thus also of LSQR) when κ is small, but poor (slow) convergence usually occurs if $\kappa \gg 1$ unless there is significant clustering of internal eigenvalues.

In Section 2, we give several examples that highlight the matrix squaring issue for linear systems and review results that indicate when it does not apply.

In section 3, we likewise consider the linear least squares problem. We also speculate why solution techniques may avoid the matrix squaring issue when $m \gg n$, but why this is less likely when n is similar to m .

We make our conclusions in section 4.

2. The matrix squaring problem.

2.1. Examples. Braess and Peisker [2] present a simple 2×2 matrix example of the matrix squaring problem. Here we present some simple examples of arbitrary dimension.

Example 1: Let B be a diagonal matrix and P be a related per-diagonal matrix:

$$B = \begin{bmatrix} b_0 & & & \\ & b_1 & & \\ & & \ddots & \\ & & & b_n \end{bmatrix}, \quad P = \begin{bmatrix} & & & b_0 \\ & & & b_1 \\ & & \ddots & \\ b_n & & & \end{bmatrix},$$

so that

$$P^{-1} = \begin{bmatrix} & & & b_n^{-1} \\ & & & b_{n-1}^{-1} \\ & & \ddots & \\ b_0^{-1} & & & \end{bmatrix}, \quad P^{-1} B = \begin{bmatrix} & & & 1 \\ & & & 1 \\ & & \ddots & \\ 1 & & & \end{bmatrix} := Y,$$

where clearly $Y^2 = I$ so $\sigma(P^{-1} B) = \{\pm 1\}$. ($P^{-1} B$ is real symmetric, hence diagonalisable). Thus for any system $Bx = b$ with any b , GMRES with preconditioner P will

converge (terminate) with the solution no later than the second iteration; P is thus an almost perfect preconditioner for B . However,

$$(P^T P)^{-1} B^T B = \begin{bmatrix} (b_0/b_n)^2 & & & \\ & (b_1/b_{n-1})^2 & & \\ & & \ddots & \\ & & & (b_n/b_0)^2 \end{bmatrix}$$

so that the eigenvalues of the preconditioned normal equations can be arbitrarily badly distributed. For example, taking $b_k = 10^k$ gives

$$\sigma((P^T P)^{-1} B^T B) = \{10^{2n-4k}, k = 0, 1, \dots, n\},$$

87 thus widely spread eigenvalues and condition number 10^{4n} . In this situation it is seen
88 that $P^T P$ is a very poor preconditioner for $B^T B$. \square

89 Though B is real symmetric in Example 1, even positive definite if all the b_i are
90 positive, it is clear that P is non-symmetric unless $b_i = b_{n-i}, i = 0, 1, \dots, n$, in which
91 case $(P^T P)^{-1} B^T B = I$. It might be considered strange to employ a non-symmetric
92 preconditioner for a symmetric matrix, but the matrix squaring problem can arise
93 even when both B and P are SPD, as our next example demonstrates.

94 *Example 2:* If $A \in \mathbb{R}^{m_1 \times n_1}$ is any matrix of full column rank $n_1 \leq m_1$, $C = QA$ for
95 any orthogonal matrix $Q \in \mathbb{R}^{m_1 \times m_1}$ and $\ell = m_1 + n_1$, then

96 (2.1)
$$B = \begin{bmatrix} 2I & C \\ C^T & 2A^T A \end{bmatrix} \in \mathbb{R}^{\ell \times \ell} \quad \text{and} \quad P = \begin{bmatrix} 2I & 0 \\ 0 & 2A^T A \end{bmatrix} \in \mathbb{R}^{\ell \times \ell}$$

are both SPD because for any non-zero $z = \begin{bmatrix} x \\ y \end{bmatrix}$, $x \in \mathbb{R}^{m_1}, y \in \mathbb{R}^{n_1}$ we have

$$z^T B z = (x + Cy)^T (x + Cy) + x^T x + y^T A^T A y > 0.$$

Further, it is readily checked that

$$\frac{1}{2} z^T P z \leq z^T B z \leq \frac{3}{2} z^T P z \quad \text{for all } z,$$

so that $\sigma(P^{-1}B) \subset [\frac{1}{2}, \frac{3}{2}]$ and we have that P is an excellent preconditioner for B
(and, because of the symmetry, P^T is an excellent preconditioner for B^T !). Since
 $\kappa \leq 3$, the CG convergence for such a problem would satisfy the convergence bound

$$\|z - z_k\|_B \leq 2 \left(\frac{\sqrt{3} - 1}{\sqrt{3} + 1} \right)^k \|z - z_0\|_B,$$

where z_k is the k^{th} iterate, z the exact solution, and $\|s\|_B^2 = s^T B s$ is the natural
norm for the problem—the norm in which the error reduces monotonically. Thus
contraction in the norm of the error of at worst 0.27 can be guaranteed at each
preconditioned CG iteration. Even if a stationary iteration for (1.1) (or (1.4)) with
preconditioner P (respectively P^T) were applied, ultimate (asymptotic) contraction
of at least $\frac{1}{2}$ would be guaranteed. It will turn out—see below—that if step-by-
step (norm) contraction of a stationary iteration for (1.4) with preconditioner P^T is
better (less) than $\sqrt{2} - 1 \approx 0.414$, then it is guaranteed that $P^T P$ must be a good

preconditioner for $B^T B$. To see in this example that $P^T P$ is arbitrarily poor as a preconditioner for $B^T B$ we simply calculate

$$\frac{z^T B^2 z}{z^T P^2 z} = \frac{4x^T x + x^T Q A A^T Q^T x}{9x^T x} \leq \lambda_{\max}(P^{-2} B^2)$$

when $z = \begin{bmatrix} x \\ 0 \end{bmatrix}$. Also

$$\frac{z^T B^2 z}{z^T P^2 z} = \frac{y^T A^T A y + 4y^T (A^T A)^2 y}{9y^T (A^T A)^2 y} \geq \lambda_{\min}(P^{-2} B^2)$$

when $z = \begin{bmatrix} 0 \\ y \end{bmatrix}$. Putting these together we have

$$\kappa = \frac{\lambda_{\max}(P^{-2} B^2)}{\lambda_{\min}(P^{-2} B^2)} \geq \frac{4 + \sigma_{\max}(A)^2}{4 + 1/\sigma_{\max}(A)^2},$$

where $\sigma_{\max}(A)$ is the largest singular value of A . This can clearly be an arbitrarily large lower bound through choice of A . The complete freedom to select A likely allows badly distributed eigenvalues between $\lambda_{\max}(P^{-2} B^2)$ and $\lambda_{\min}(P^{-2} B^2)$.

Note that we are not saying that there might not be more preferable ways to solve linear systems involving B in (2.1), but a block-diagonal approximation/preconditioner such as P in (2.1) is not so unreasonable.

□

Symmetry is certainly not required to cause the matrix squaring problem. Indeed, without any preconditioning, Nachtigal, Reddy and Trefethen [11] years ago identified examples of $n \times n$ matrices where LSQR takes $n/2$ times more iterations than GMRES for convergence and vice versa. By simply writing such examples as the product of a matrix B and appropriate P^{-1} one obviously has the same result!

So for our third example we turn to something more akin to a practical problem to show how the matrix squaring issue varies as we vary problem parameters. We employ a simple but reasonable discretisation of a differential equation problem and a heuristically reasonable preconditioner. The resulting matrix is tridiagonal, so this is certainly not a challenging problem of linear algebra.

Example 3: We consider the one-dimensional convection-diffusion problem

$$-\nu \frac{d^2 u}{dx^2} + \frac{du}{dx} = 0, \quad x \in (0, 1)$$

with boundary conditions $u(0) = 1, u(1) = 0$. For any small $\nu > 0$ the solution contains a boundary layer of thickness $O(\nu)$ near 1 where it reduces exponentially from approximately 1 to 0. A central-difference approximation

$$-\nu(u_{j+1} - 2u_j + u_{j-1})/h^2 + (u_{j+1} - u_{j-1})/2h = 0$$

yields a tridiagonal system $Bu = b$, where

$$B = \text{tridiag}\left(-\frac{\nu}{h^2} - \frac{1}{2h}, \frac{2\nu}{h^2}, -\frac{\nu}{h^2} + \frac{1}{2h}\right),$$

and b is the zero vector apart from its first entry, $\nu/h^2 + 1/2h$. For small values of ν the differential equation is dominated by the convection, so we employ an upwind

(first-order) difference for the first derivative as a preconditioner, yielding a lower bidiagonal matrix

$$P = \frac{1}{h} \begin{bmatrix} \ddots & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & -1 & 1 \\ & & & & \ddots & \ddots \end{bmatrix}.$$

In Table 1 we display the number of GMRES and LSQR iterations to achieve the given tolerance for a range of values of ν . In all cases the matrix dimension is 1000. In each case the solution accuracy is acceptable given the tolerance; it is quite easy for this problem to specify parameter values for which the iterative solution remains far from the actual solution, particularly with LSQR. Without preconditioning, both iterative methods are generally poor for this problem, LSQR particularly so. Applying CG to the normal equations is apparently significantly worse than LSQR.

TABLE 1
Iteration counts for example 3.

ν	tolerance	GMRES	LSQR	$\ I - BP^{-1}\ _2$
0.001	1.e-3	4	14	15.87
0.001	1.e-6	10	21	15.87
0.005	1.e-3	15	66	142.7
0.005	1.e-6	49	108	142.7
0.01	1.e-6	95	198	301.2
0.05	1.e-6	434	751	1569.3

Note that iterations rather than computational work are the results presented here: it is well known that the computational work for GMRES increases with each iteration whereas LSQR has fixed work per iteration. It is apparent that the preconditioner is less effective for larger values of ν as is to be expected, and that the extreme matrix squaring issue in example 2 is far less acute here.

The MATLAB functions `gmres` and `lsqr` were used for these simple computations.

2.2. Theory. The diversity of the above examples indicates that the matrix squaring problem is far from unusual. However, there are situations where the problem can not arise.

THEOREM 2.1. *If $\|I - BP^{-1}\| < \sqrt{2} - 1$ then $\sigma((P^T P)^{-1} B^T B) \subset (0, 2)$. Moreover if $\|I - BP^{-1}\| = \sqrt{2} - 1 - \delta$ then*

$$\lambda_{\min}((P^T P)^{-1} B^T B) \geq \sqrt{2}\delta + \delta^2, \quad \lambda_{\max}((P^T P)^{-1} B^T B) \leq 2 - \sqrt{2}\delta - \delta^2.$$

Proof. Let $T = I - BP^{-1}$, so $T^T = I - P^{-T} B^T$. Note that $(P^T P)^{-1} B^T B = P^{-1} P^{-T} B^T B$ is similar to the SPD matrix

$$P^{-T} B^T B P^{-1} = (I - T^T)(I - T) = I - T - T^T + T^T T.$$

The eigenvalues of this matrix are real and satisfy

$$(2.2) \quad 1 - 2\|T\| - \|T\|^2 \leq \lambda \leq 1 + 2\|T\| + \|T\|^2,$$

because $\|T^T\| = \|T\|$ and thus $\|T^T T\| \leq \|T\|^2$. Now $2r + r^2 < 1$ for positive r when $r < \sqrt{2} - 1$ and the first result follows.

Setting $\|T\| = \|I - BP^{-1}\| = \sqrt{2} - 1 - \delta$ in (2.2) yields the final statement. \square

Thus if the stationary iteration (1.5) for (1.4) is sufficiently contractive, Theorem 2.1 provides bounds on the condition number of the preconditioned normal equations. The smaller is the contraction $\|I - BP^{-1}\| < \sqrt{2} - 1$, the better $P^T P$ is as a preconditioner for $B^T B$. In a less quantitative form than in the theorem here, this was already noted by Braess and Peisker.

Remark 2.2. If the condition of Theorem 2.1 is satisfied, we have a bound for the condition number κ of the preconditioned normal equations:

$$\kappa \leq (2 - \sqrt{2}\delta - \delta^2) / (\sqrt{2}\delta + \delta^2)$$

and thus an upper bound on CG (and LSQR) convergence via (1.6). For example, if $\delta = 0.1$, then CG (and LSQR) iteration error must contract by 0.56 or better at every iteration in the natural norm. If $\delta = 0.01$, then contraction per iteration must be by 0.85 or better in the natural norm.

The above theorem is essentially given in the recent meteorological literature by Gratton et al. [7], whose primary concern is the weighted linear least squares problem that arises in data assimilation.

A partial (and perhaps less useful) converse is also possible in the case that B and P are both SPD.

THEOREM 2.3. If $B, P \in \mathbb{R}^{m \times m}$ are both SPD and

$$\lambda_{\min}(B + P^{-1}) \leq \epsilon, \quad \lambda_{\max}(BP^{-1} + P^{-T}B^T) \geq \Upsilon > 0,$$

then $\kappa((P^T P)^{-1}B^T B) \geq 4\Upsilon^2/\epsilon^4$.

Proof. First P^{-1} and thus $B + P^{-1}$ must also be SPD. We directly apply Theorem 1 in [3] to obtain

$$\sqrt{\sigma_{\min}(BP^{-1})} \leq \frac{1}{2}\lambda_{\min}(B + P^{-1}) \leq \frac{1}{2}\epsilon.$$

Thus

$$(2.3) \quad \lambda_{\min}((P^T P)^{-1}B^T B) = \lambda_{\min}((BP^{-1})^T BP^{-1}) \leq \epsilon^4/16$$

by taking the fourth power, as the eigenvalues here are the squares of the singular values of BP^{-1} and we have used the similarity transform

$$P((P^T P)^{-1}B^T B)P^{-1} = (BP^{-1})^T BP^{-1}.$$

For a lower bound on the largest singular value, using [9, page 151] we have

$$\sigma_{\max}(BP^{-1}) \geq \lambda_{\max}\left(\frac{BP^{-1} + P^{-T}B^T}{2}\right) \geq \Upsilon/2,$$

and squaring gives

$$(2.4) \quad \lambda_{\max}((P^T P)^{-1}B^T B) = \lambda_{\max}((BP^{-1})^T BP^{-1}) \geq \Upsilon^2/4.$$

Taking the quotient of (2.4) with (2.3) gives the stated result. \square

Note that with the assumptions in Theorem 2.3 we have

$$\kappa(BP^{-1}) = \frac{\sigma_{\max}(BP^{-1})}{\sigma_{\min}(BP^{-1})} \geq \frac{2\Upsilon}{\epsilon^2}$$

(which does not imply that either $\|BP^{-1}\|$ or $\|I - BP^{-1}\|$ is necessarily large). Thus the condition number of BP^{-1} and of the preconditioned normal equations must be large (and slow convergence of CG and LSQR is the likely consequence) if $B + P^{-1}$ has small eigenvalues and the symmetric part of BP^{-1} has large eigenvalues.

3. Preconditioned linear least squares. The solution of large linear least squares problems almost always relies on the use of LSQR (see, for example [10]). With $B \in \mathbb{R}^{m \times n}$ ($m \geq n$) and a (right) preconditioner $R \in \mathbb{R}^{n \times n}$, the solution of

$$y = \operatorname{argmin} \|b - BR^{-1}y\|, \quad x = R^{-1}y$$

by LSQR is mathematically equivalent to applying CG to the preconditioned normal equations

$$(3.1) \quad R^{-T}B^TBR^{-1}y = R^{-T}B^Tb.$$

Some authors do use CG in their computations. The fact that LSQR is applicable to determined or overdetermined systems is very useful and it is usually more numerically stable and needs fewer iterations.

It is clear that if R can be selected such that $\kappa(BR^{-1}) = \sigma_{\max}(BR^{-1})/\sigma_{\min}(BR^{-1})$ is small, then this necessarily leads to rapid iterative convergence to the solution because the squares of the singular values are the eigenvalues of the preconditioned normal matrix $(BR^{-1})^TBR^{-1} = R^{-T}B^TBR^{-1}$. For example, in consideration of a randomized algorithm for the generation of R when $m \gg n$, Rokhlin and Tygert [14] establish small upper bounds on $\kappa(BR^{-1})$ that hold with high probability; they are then able to use R as a reliable preconditioner with CG for the solution of (3.1). This consideration of singular values rather than eigenvalues is precisely the right thing to do; [14] has provided the springboard for further powerful and well-founded methods for least squares problems when $m \gg n$, though their emphasis is not on large and sparse problems. Equally clearly, when $m = n$ the considerations in the previous section directly apply, and the matrix squaring problem could arise and cause slow convergence.

A simple example indicates that there can be matrix squaring type difficulties also in the case of least squares:

Example 4: If

$$B = \begin{bmatrix} C \\ D \end{bmatrix} \in \mathbb{R}^{2n \times n},$$

where C, D are both square and $C = QR$ with Q orthogonal, then

$$BR^{-1} = \begin{bmatrix} Q \\ DR^{-1} \end{bmatrix}$$

and

$$(3.2) \quad R^{-T}B^TBR^{-1} = I + R^{-T}D^TDR^{-1}.$$

Now, if D is any one of the square matrices denoted by B in Examples 1, 2 or 3 and R is the square matrix denoted by P in the relevant example, then the eigenvalues of DR^{-1} can be nicely distributed (clustered) whereas those of $R^{-T}D^TDR^{-1}$ can be poorly distributed (widely spread). Thus, even though the smallest eigenvalue here can be no less than 1, the largest eigenvalue of (3.2) can be large so that CG for the corresponding normal equations—and so likewise LSQR for the least squares problem—could converge slowly. Note that by identifying D and R , we are essentially limiting C to be an orthogonal transform of R . Note also that there is no stipulation that R be triangular here, though in practical situations, that is often the case.

□

It is clear that if $n \ll m$, the smaller dimension of the normal equations helps: the theoretical maximum number of LSQR iterations is n regardless of any preconditioning. However, if m and n are of more comparable size, then the effectiveness of LSQR will depend on any preconditioning required, and then the matrix squaring issue can arise.

4. Conclusions. For linear systems of equations, the derivation of preconditioners for the normal equations can be difficult if the matrix squaring problem arises. Indeed, the more widespread use in applications of non-symmetric iterative methods such as GMRES, rather than symmetric methods applied to the normal equations, may be precisely because of the issue of identifying effective preconditioners. We have argued that even with a good preconditioner for a given matrix, it does not follow that this helps in identifying a good preconditioner for the normal equations.

For the linear least squares problem, when iterative methods are needed because matrix factorizations are not practical, the matrix squaring problem remains an important consideration. As such, it seems generally more difficult to identify preconditioners for least squares problems.

Acknowledgments. I sincerely thank Michael Saunders for the several times he has asked me why people more often use GMRES than LSQR for solving square linear systems. I hope to have provided a possible answer here.

I am also grateful to three anonymous referees, whose comments have clarified and greatly improved this short manuscript.

REFERENCES

- [1] M. BENZI, *Preconditioning techniques for large linear systems: A survey*, Journal of Computational Physics, 182 (2002), pp. 418–477, <https://doi.org/10.1006/jcph.2002.7176>.
- [2] D. BRAESS AND P. PEISKER, *On the numerical solution of the biharmonic equation and the role of squaring matrices for preconditioning*, IMA Journal of Numerical Analysis, 6 (1986), pp. 393–404, <https://doi.org/10.1093/imanum/6.4.393>.
- [3] S. DRURY, *On a question of Bhatia and Kittaneh*, Linear Algebra and its Applications, 437 (2012), pp. 1955–1960, <https://doi.org/10.1016/j.laa.2012.04.040>.
- [4] H. ELMAN, D. SILVESTER, AND A. WATHEN, *Finite Elements and Fast Iterative Solvers: With Applications in Incompressible Fluid Dynamics*, Oxford University Press, United Kingdom, second ed., 2014.
- [5] M. FERRONATO, *Preconditioning for sparse linear systems at the dawn of the 21st century: history, current developments, and future perspectives*, International Scholarly Research Notices, 2012, Article ID 127647, 49 pages (2012), <https://doi.org/10.5402/2012/127647>.
- [6] R. FREUND AND N. NACHTIGAL, *QMR: a quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 315–339, <https://doi.org/10.1007/BF01385726>.
- [7] S. GRATTON, S. GÜROL, E. SIMON, AND P. L. TOINT, *A note on preconditioning weighted linear least-squares, with consequences for weakly constrained variational data assimilation*, Quarterly Journal of the Royal Meteorological Society, 144 (2018), pp. 934–940, <https://doi.org/10.1002/qj.3262>.
- [8] M. R. HESTENES AND E. STIEFEL, *Methods of Conjugate Gradients for solving linear systems*, J Res NIST, 49 (1952), pp. 409–436, <https://doi.org/10.6028/jres.049.044>.
- [9] R. A. HORN AND C. R. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge; New York, 1994.
- [10] X. MENG, M. A. SAUNDERS, AND M. W. MAHONEY, *LSRN: a parallel iterative solver for strongly over- or underdetermined systems*, SIAM J. Sci. Comput., 36 (2014), pp. C95–C118.
- [11] N. M. NACHTIGAL, S. C. REDDY, AND L. TREFETHEN, *How fast are nonsymmetric matrix iterations?*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 778–795, <https://doi.org/10.1137/0613049>.
- [12] C. C. PAIGE AND M. A. SAUNDERS, *Solution of sparse indefinite systems of linear equations*, SIAM Journal on Numerical Analysis, 12 (1975), pp. 617–629, <http://www.jstor>.

- org/stable/2156178.
- [13] C. C. PAIGE AND M. A. SAUNDERS, *LSQR: An algorithm for sparse linear equations and sparse least squares*, ACM Trans. Math. Software, 8 (1982), pp. 43–71, <https://doi.org/10.1145/355984.355989>.
 - [14] V. ROKHLIN AND M. TYGERT, *A fast randomized algorithm for overdetermined linear least-squares regression*, Proceedings of the National Academy of Sciences, 105 (2008), pp. 13212–13217, <https://doi.org/10.1073/pnas.0804869105>.
 - [15] Y. SAAD, *Iterative Methods for Sparse Linear Systems*, SIAM, second ed., 2003, <https://doi.org/10.1137/1.9780898718003>.
 - [16] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on Scientific and Statistical Computing, 7 (1986), pp. 856–869, <https://doi.org/10.1137/0907058>.
 - [17] J. SCOTT AND M. TUMA, *Preconditioning of linear least squares by robust incomplete factorization for implicitly held normal equations*, SIAM Journal on Scientific Computing, 38 (2016), pp. C603–C623.
 - [18] P. SONNEVELD AND M. VAN GIJZEN, *IDR(s): a family of simple and fast algorithms for solving large nonsymmetric linear systems*, SIAM Journal on Scientific Computing, 31 (2008), pp. 1035–1062, <https://doi.org/10.1137/070685804>.
 - [19] H. A. VAN DER VORST, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems.*, SIAM J. Sci. Comput., 13 (1992), pp. 631–644, <https://doi.org/10.1137/0913035>.
 - [20] H. A. VAN DER VORST, *Iterative Krylov Methods for Large Linear Systems*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, 2003, <https://doi.org/10.1017/CBO9780511615115>.
 - [21] A. J. WATHEN, *Preconditioning*, Acta Numerica, 24 (2015), pp. 329–376, <https://doi.org/10.1017/S0962492915000021>.